"How short is a piece of string?" The Impact of Text Length and Text Augmentation on Short-text Classification Accuracy

Austin McCartney, Svetlana Hensman, Luca Longo

School of Computing, Dublin Institute of Technology, Dublin, Ireland
austin.mccartney@gmail.com svetlana.hensman@dit.ie luca.longo@dit.ie

Abstract. Recent increases in the use and availability of short messages have created opportunities to harvest vast amounts of information through machine-based classification. However, traditional classification methods have failed to yield accuracies comparable to classification accuracies on longer texts. Several approaches have previously been employed to extend traditional methods to overcome this problem, including the enhancement of the original texts through the construction of associations with external data supplementation sources. Existing literature does not precisely describe the impact of text length on classification performance. This work quantitatively examines the changes in accuracy of a small selection of classifiers using a variety of enhancement methods, as text length progressively decreases. Findings, based on ANOVA testing at a 95% confidence interval, suggest that the performance of classifiers using simple enhancements decreases with decreasing text length, but that the use of more sophisticated enhancements risks over-supplementation of the text and consequent concept drift and classification performance decrease as text length increases.

1 Introduction

Traditional techniques for machine classification of texts rely on statistical methods which, in turn, rely on a sufficiency of meaningful data (words) within the texts to allow classification. In the case of short texts, the performance of such classifiers is reported as being poor in comparison with performance on longer texts, the inference being that insufficient data is present within the target texts. One approach to the improvement of classifier performance has been the augmentation or enhancement of the short text by the addition of synonyms, or other semantically linked words, to the body of the original text prior to classification. The implicit hope in such supplementation is that the additional words are conceptually related to the words in the original text and will therefore amplify the underlying meaning and context of the original. Despite quite extensive coverage in published literature of the general area of short test classification, very little specific information has been available relating to the deterioration of

classifier performance on shorter texts; the exact nature of the relationship between text length and classifier performance has been unclear and, consequently, no common definition of how short a target text may be before it can be considered troublesome is available. An attempt will be made to address the question of how text-length, message enhancement and accuracy interact, through the repeated classification of enhanced texts of controlled lengths. Three common classifiers will be used to rule out the possibility of results specific to a single classifier.

The remainder of this paper will be laid out as follows: Section 2 will review the published literature relating to relevant, similar, work. Section 3 will discuss the design and execution of the experiments used in this study. Section 4 will present the results of experiments and further statistical analysis. Section 5 will close with conclusions and suggestions for future related work.

2 Related Work

A variety of different techniques have been proposed to enhance or enrich short texts by the addition of extra features designed to make matching, clustering and classification easier. Some of these methods rely on the exploitation of external taxonomies, typically Wikipedia or Probase, whereas others use semantic nets such as Wordnet. Song, Ye, Du, Huang and Bie [19] present a survey of short text classification, first giving an overview of the special conditions which attach to short text as a problem, and then outlining all the major avenues of current research. They divide approaches into three broad families; semantic approaches (including LSA), semi-supervised classical methods (e.g. SVM, naïve-Bayes) and ensemble methods, which can combine from the other two families.

Work presented by Bollegala, Matsuo and Ishizuka [3] incorporates semantic information extracted from web-based search engines and this is contrasted with the same operation using Wordnet: the authors point out that, typically, a static resource such as Wordnet will fail to produce good results when trying to judge similarity in the presence of colloquialisms. This use of an explicit external taxonomy such as Wordnet can be contrasted with much work which makes use of the implicit taxonomy inherent in the organisation and content of reference sources such as Wikipedia and Probase, as in the work of Banerjee, Ramanathan, and Gupta [1], where the titles of Wikipedia articles containing terms of interest are used as features to supplement the sparse text data, or in the work of Wang, Wang, Li, and Wen [21] in which they coin the term "bag-of-concepts" to stress the semantic aspect of the additional features that they had mined from the probabilistic semantic network Probase. Wikipedia is once again the favoured external source of "world knowledge" in Gabrilovich and Markovitch [7] in which they state, "pruning the inverted index (concept selection) is vital in eliminating noise", but, unfortunately, they provide no further detail on their "ablation" process. Gabrilovich and Markovitch go on to claim double digit im-

provements over the then state-of-the-art methods on certain datasets. Genc, Sakamoto and Nickerson [9] compare three disparate techniques to demonstrate the utility of Wikipedia as an implicit taxonomic source. In a manner similar to, but subtly different from, Gabrilovich and Markovitch [8] they use the target text to mine relevant Wikipedia pages, and then calculate the distances between Wikipedia pages using a simple shortest path graph traversal metric to assign distances between target texts. Their second technique is to simply measure the String Edit Distance, (SED), between texts using the Levenshtein metric. Their final design uses Latent Semantic Analysis, (LSA), coupled with a cosine distance metric. Their results suggest that the Wikipedia method out-performed both SED and LSA on most sets, and was inferior on none of the tested datasets.

Departing from the common themes above, Sun [20] takes a distinctly different direction to the main approaches outlined above, and trims short texts even further in an attempt to retain only key words. Trimming is accomplished using familiar term-frequency / inverse-document-frequency methods coupled with a novel clarity measure, and is followed with a classification implemented through a Lucene search to find similar documents from a corpus: the classes of the returned documents are used as the class for document under classification. Sun reports that results match MaxEnt classifiers. A trend in the short text enhancement literature becomes apparent over time: early work concentrated on well-structured external resources such as Wordnet but, with time, the favoured approach became the more unstructured Wikipedia-type model. Although frequent reference is made to the difficulty of classifying short text, as for example in Song, Ye, Du, Huang and Bie [19], all bar one of the reviewed articles omit any reference to the quantitative impact of the shortness of the text or any definition of how short a text must be to be considered "short". Yuan, Cong and Thalmann [23] in their paper, which is concerned, primarily, with contrasting various smoothing methods as applied to naïve-Bayes, conclude only that classifiers perform more poorly with single word texts than with multi-word texts. It is this gap in existing research which underpins the motivation for the current work.

3 Methods

The fundamental design of this project's experimental work centres on measuring binary classification performance on enhanced variants of messages of known specific lengths dependent on message contents having either positive or negative sentiment. The decision to choose binary sentiment classification as the reference task was motivated by the the fact that although it represents a real-world application it remains relatively free of additional complexity that might complicate analysis of results. The differences in classification performance of three common classifiers, across message lengths and across enhancement methods, as measured by the F1 score for accuracy of classification, were analysed to determine if message length or enhancement has any statistically valid impact

on classification performance.The experimental data was a corpus of 1.8 million pre-classified and pre-cleaned micro-blog (twitter) posts of all lengths obtained from the Sentiment140[1] sentiment analysis project run by Stanford University and described by Go, Bhayani and Huang [10].

3.1 Data Preparation

The original data set from the Sentiment140 project was split into subsets by exact message-length, each subset containing 5000 tweets, all of exactly the same length and having an even balance between tweets having positive and negative sentiment. There were twelve length categories, as measured by the total number of characters in the original message, as follows: 138, 110, 80, 50, 45, 40, 35, 30, 25, 20, 15 characters, and a final set of tweets of length ≤ 10 characters.

3.2 Data Pre-processing

Each tweet message in each of the length-determined subsets was pre-treated with nine text enhancement techniques to produce a total of ten variants of each message, including the original message. Three approaches to enhancement were used: basic, Wordnet-based and Wikipedia-based.

Basic Enhancements Basic enhancements consist of operations such as the removal of stop words, punctuation and twitter hashtags, the lemmatization of the text and the creation of bigrams. Specific basic enhancements were:

- **Original** - the original text of the tweet from the Sentiment140 dataset.
- **Cleaned** - the original text having punctuation and stop words removed, and twitter specific strings (e.g. hashtags, URLs) replaced with standard tokens.
- **Lemmatised** - the cleaned set (above) lemmatised using the NLTK python library.
- **Bigrams** - Appending all bigrams from the lemmatised tweet back to the lemmatized tweet.

Wordnet Wordnet [17] is a semantically focused English language dictionary. It bears a resemblance to an extended thesaurus but, importantly from the perspective of this work, it contains not only synonyms, but also hypernyms and hyponyms. Specific Wordnet enhancements were:

- **Synonyms** - enhanced by appending all available wordnet synonyms for each word in the lemmatised tweet to the lemmatized tweet.
- **Hypernyms** - enhanced by appending all available wordnet hypernyms for each word in the lemmatised tweet back to the lemmatized tweet.
- **Hyponyms** - enhanced by appending all available wordnet hyponyms for each word in the lemmatised tweet back to the lemmatized tweet.

[1] `http://help.sentiment140.com/for-students/`

Wikipedia / DBpedia DBpedia is a static, structured, database derived from information contained in the on-line encyclopaedia Wikipedia. DBpedia returns, in XML format, the Wikipedia taxonomic metadata for the most relevant Wikipedia pages when a given word or bigram is searched. This metadata includes page titles, Wikipedia categories and Wikipedia classes. These metadata each have a "label" which is a text descriptor, possibly containing multiple words, of the page title, category or class. Specific Wikipedia enhancements were:

- **Wiki Words** - enhanced by appending all available words in all the labels contained in the top five Wikipedia hits for each word in the lemmatised text back to the lemmatised text.
- **Wiki Phrases** - enhanced by appending all available labels, each treated as an indivisible string (n-gram), from the top five Wikipedia hits for each word in the lemmatised text back to the lemmatised text.
- **Wiki Bigrams** - enhanced by appending all available labels, each treated as an indivisible string (n-gram), from the top five Wikipedia hits for each bigram in the lemmatised text back to the lemmatised text.

It may be noted that these three approaches to enhancement can be categorised into one of two classes: the basic enhancements do not supplement the text with any external data if we discount the substitution of a word with its own lemma, and so they can be considered "non-additive", whereas the Wordnet and Wikipedia/DBpedia approaches rely primarily on the addition of external data which, it is implicitly hoped, is in some way conceptually linked to the words in the original text, thereby amplifying the underlying meaning of the text. The latter methods may be considered "additive".

3.3 Modelling

The three common classifiers used in the experiment were:

- **Naïve-Bayes** [14], [18]
- **Support Vector Machine (SVM)** [5], [13]
- **Latent Semantic Analysis (LSA)** [6], [15]

No attempt, beyond the most basic, was made to optimise or tune classifier performance and any reference to the comparative performance of classifiers is made in an informal sense. The use of multiple classifiers was undertaken only in order to demonstrate the general applicability of the findings, if any, and to rule out any effect that may arise from the use of any specific classifier: reflecting this purpose, the three classifiers chosen were used in their most basic configurations and used the built-in routines from the scikit-learn python library. Each of the 120 resultant data sets of 5000 tweets (10 enhancements for each of 12 text lengths) was classified by each of the three classifiers after 100 repetitions of a Monte Carlo cross-validation using a 90% training and 10% test split of the data. The mean F1 score for classification accuracy was calculated for each of the 100-fold cross validations. This eventually yielded three results sets of F1 accuracy scores, one for each classifier, each containing an average F1 score for each of the 120 combinations of enhancement and text length.

3.4 Evaluation

The sets of mean F1 Scores for each classifier-enhancement combination were subjected to Wilcox's trimmed means robust 1-way ANOVA testing [22], at the 95% confidence level, to determine if text length had a significant impact on classification accuracy. This was followed by Wilcox's robust 2-way ANOVA testing, at the 95% confidence level, on each classifier's data set to determine whether there was a statistically significant interaction between text length and enhancement method which influenced accuracy. An approximate measure of the overall accuracy of each enhancement-classifier combination was made by summing the accuracy results for all text lengths for each combination - this may be thought of as a crude measure of the area-under-the-curve for plots of accuracy (y-axis) drawn on a text-length abscissa (x-axis). The enhanced data sets were analysed to calculate the average relative size of their texts compared to the original texts. For example, if the mean length of synonym-enhancements for original messages of length 20 characters was found to be 140 characters, the "additive footprint" for synonym enhancement at 20 characters would be calculated to be 7.0. Additive footprint for a given enhancement was found, by ANOVA, not to vary significantly as a function of text length and so may be thought of as characteristic of an enhancement as a whole. Both additive footprints and overall accuracy for each classifier-enhancement combination were rank-ordered, and Spearman's Rank-Order co-efficient test was carried out to determine whether the additive footprint of an enhancement was correlated with the overall classification accuracy of that enhancement for a given classifier.

4 Results

Numeric accuracy results for all three classifiers are omitted in the interest of brevity. Instead, accuracy results in graphical form are presented along with tabular results for additive footprint calculations and rank correlation results. 1-Way robust ANOVA conducted on each enhancement for each classifier indicates that significant (95%, $p \leq 0.0001$) differences are present as text length changes for all combinations. This finding supports rejection of the hypothesis that text length does not influence classification accuracy. 2-way robust ANOVA across text-lengths and enhancements within each classifier indicates that a significant interaction (95%, $p \leq 0.001$) exists between text length and enhancement for all classifiers. This finding supports rejection of the hypothesis that the chosen enhancement method has no significant effect on the way in which the F1 score changes with changes in text length. Note that, on all three sets of classifier plots, local or absolute maxima for accuracy are frequently observed in text-lengths from 20 to 25 characters. The mean additive footprint of each enhancement and the "area under the accuracy curve" for each enhancement-classifier combination were calculated. These tabular results are displayed in addition to the graphical accuracy results.

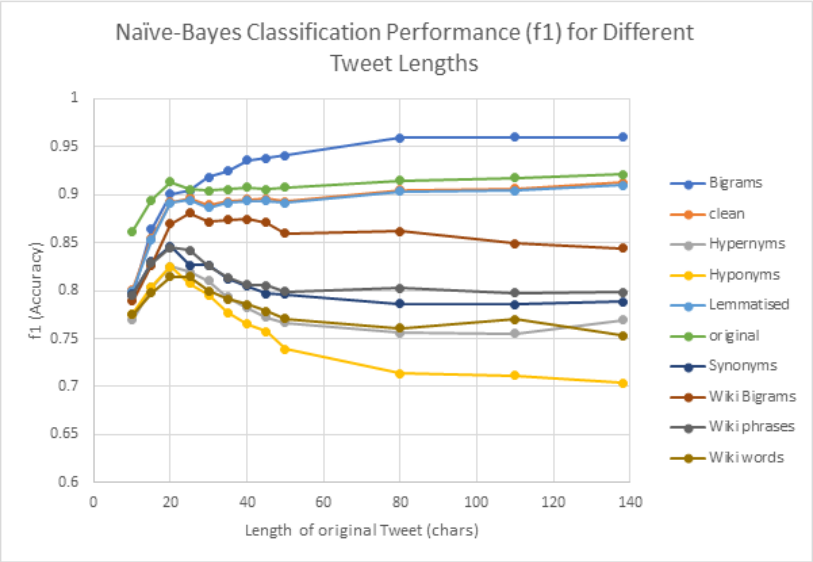

Fig. 1. Plots of Accuracy vs. Text Length for All Enhancements using Naïve-Bayes

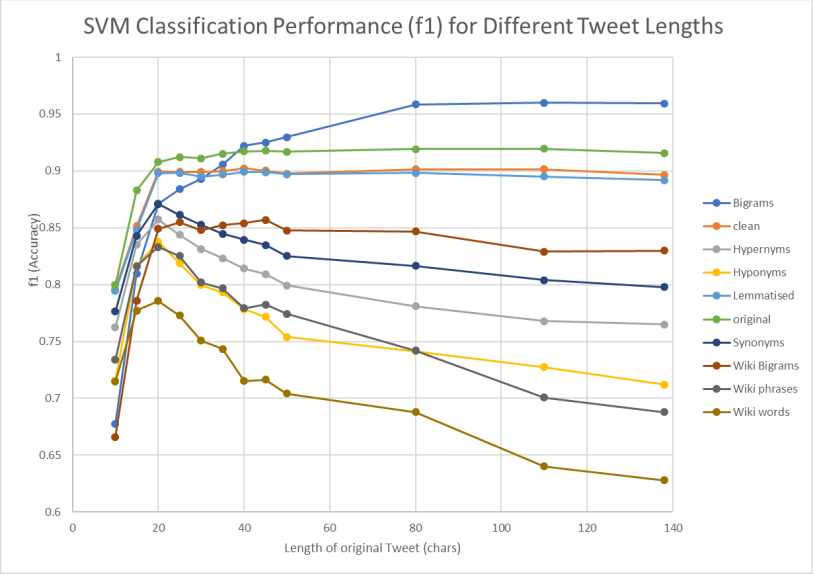

Fig. 2. Plots of Accuracy vs. Text Length for All Enhancements using SVM

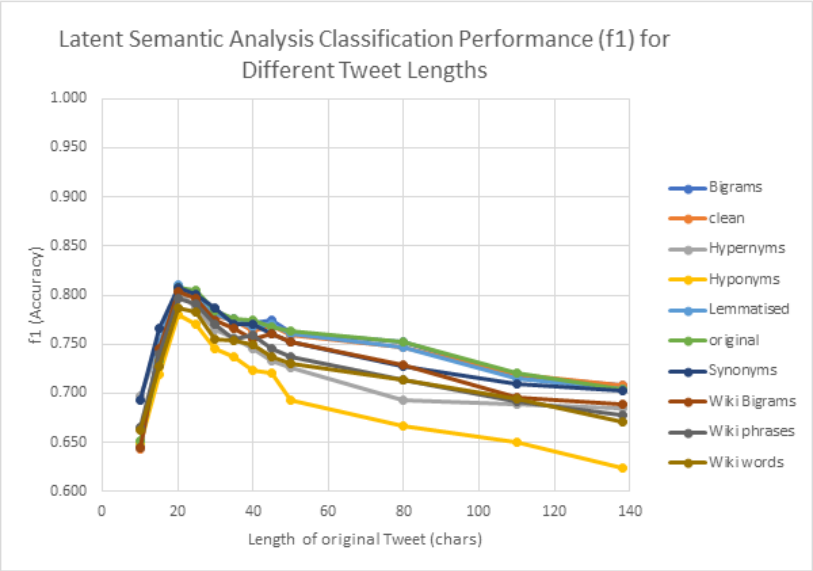

Fig. 3. Plots of Accuracy vs. Text Length for All Enhancements using LSA

Table 1 shows the mean additive footprint calculated for each enhancement, along with the relative rank of the footprint size. Table 2 gives the nominal area under each of the accuracy curves, obtained by adding the point values for each curve. Table 3 shows the ranked footprints from Table 1, alongside the ranked score for the nominal area under the accuracy curve for each enhancement within each classifier from Table 2. These ranked summation figures represent the comparative overall accuracy of a given enhancement for each classifier.

Table 1. Additive Footprint Scores and Relative Rankings for All Enhancements

AF Rank	Enhancement									
Footprints	Original	Clean	Lemmatised	Wiki words	Wiki phrases	Synonyms	Hyponyms	Hypernyms	Bigrams	Wiki Bigrams
Footprint	**1**	**0.7**	**0.7**	**56.1**	**20.5**	**7.8**	**19.8**	**7.7**	**0.9**	**1.9**
Rank	4	1.5	1.5	10	9	7	8	6	3	5

Table 2. Areas under the Accuracy Curve for Classifiers and Enhancements

AUC	Enhancement									
Classifier	Bigrams	Clean	Hypernyms	Hyponyms	Lemmatised	Original	Synonyms	Wiki Bigrams	Wiki phrases	Wiki words
NB	11	10.632	9.421	9.176	10.61	10.859	9.695	10.271	9.761	9.411
SVM	10.696	10.645	9.69	9.266	10.611	10.836	9.967	9.92	9.274	8.636
LSA	9.047	9.017	8.828	8.479	9.024	9.038	9.048	8.914	8.845	8.764

Table 3. Correlation between Relative Accuracy Rankings and Additive Footprint Rankings for Classifiers and Enhancements

Ranks	Enhancement											
Classifier	Bigrams	Clean	Hypernyms	Hyponyms	Lemmatised	Original	Synonyms	Wiki Bigrams	Wiki phrases	Wiki words	Spearman's r	z-score
NB	1	3	8	10	4	2	7	5	6	9	0.790	2.37
SVM	2	3	7	9	4	1	5	6	8	10	0.839	2.52
LSA	2	5	8	10	4	3	1	6	7	9	0.571	1.71
Footprint	3	1.5	6	8	1.5	4	7	5	9	10		

The values of Spearman's test indicate a strong correlation between increasing additive footprint and decreasing accuracy as measured by F1 score for the naïve-Bayes and SVM classifiers, and a moderate correlation for the LSA classifier. In all three cases, the one-tailed z-score indicates a significant correlation between increasing additive footprint and decreasing accuracy at the 95% confidence level.

This empirical result would suggest that enhancements which over-supplement the original text are likely to be counter-productive in terms of accurate classification, and that the greater the degree of over-supplementation the greater the negative impact on classification accuracy. Visual inspection of the graphical data shows that not only do additive enhancements under-perform non-additive enhancements in this experiment, but that they also actually decrease classification performance as the text length increases. It is postulated that additive enhancement methods, without careful control, may overwhelm any actual signal present in the text though the addition of noise associated with poorly matched textual supplementation and that the associated concept drift will decrease clas-

sification accuracy. Qualitative changes in accuracy can be seen to start as text length decreases towards 50 characters for all non-additive enhancements, and become very pronounced below 20 characters for all variants of a message. This intuitive analysis was supported by post-hoc testing which also indicated that, for stable enhancements, statistically significant changes started to occur below 80 characters. This in turn suggests that, if the cases of naïve-Bayes and SVM classifiers can be taken to be representative, text might be usefully, if subjectively, considered short at lengths below 80 characters and very short at lengths of less than 20 characters. The LSA classifier shows a decrease in accuracy across all enhancements with increasing length beyond 25 characters: both this behaviour, and the root cause of the comparative under-performance of the LSA classifier, remain open issues for further investigation, but it should be noted that the unsupervised nature of the LSA classifier might reasonably be expected to perform less well than the supervised tasks on this particular problem. In contrast, while SVM has been recognised as a strong performer, several authors explicitly suggest that naïve-Bayes is often under-estimated [16] and, given large, balanced datasets and consistent document lengths, as in this case, may perform on a par with more sophisticated algorithms [14] [18] [4]. In a more general sense, Holte [12] observes that simple problems often respond very well to simple classification approaches and both Halevy, Norvig and Pereira [11] and Banko and Brill [2] emphasise the importance of data characteristics over specific algorithm choice. Against this backdrop, the relative strength of the naïve-Bayes classifier in this experiment should not be considered anomalous.

5 Conclusion

Addressing a lack of quantitative experimental work on the often-discussed impact of text-length upon classification accuracy, this work undertook to investigate the relationship between text-length, textual enhancement and classification accuracy by means of an experiment in which messages of carefully controlled length were enhanced using variations on common text supplementation methods and were then repeatedly classified. The primary contribution of this work is to have provided direct, quantitative, experimental, evidence that classification accuracy, for two of three tested classifiers, declines with declining text length for non-additive text enhancements, and that the exact quantitative nature of that decline was dependent upon the enhancement or pre-treatment applied to the text and to the classifier in use. The concept of "additive footprint" was introduced to quantify the proportional increase in word count imposed upon a text by a given enhancement, and it was found that the additive footprint remained, for this data set, relatively constant for a given enhancement over a range of text lengths and can thus be considered characteristic of an enhancement method, independent of text length. The findings related to additive enhancements may seem, at first glance, to contradict many published successes in the area of short text enhancement. However, the particular difficulties encountered in the supplementation of short text have been obliquely alluded to by several authors [7] [20].

The salient finding is that, without some form filtering, textual supplementation, has proved to be worse than useless. It is perhaps instructive to note that at the very shortest text lengths, the highest performing 'enhancement' was the original message which was completely un-enhanced.

Future work might usefully investigate the "bump" in accuracy seen for many enhancement-classifier combinations at message lengths of 20 to 25 characters. Some preliminary investigation was conducted to rule out any peculiarity or data artefact that may cause this small increase in accuracy, but replacement of the original data sets had no effect. A carefully designed experiment may be able to determine whether author-created context and structure inherently varies with text length: for example, it may indicate that texts in the 20 to 25-character range have a higher degree of author-created clarity, which might, tentatively, be attributed to an author's avoidance of ambiguity when composing shorter messages. Another possible avenue for future work on additive enhancement methods is experimentation with part-of-speech filtering, either at generation time (e.g. send only adjectives to wordnet for supplementation) or at application time (e.g. accept only adjectives as supplemental words) or both together. Such a filtering mechanism could be potentially used to attempt to limit the addition of non-relevant words to the original text. The narrow experimental focus of the experimental work described, in terms of classifiers, enhancements, classification task and datasets provides ample opportunity for the further exploration of the generalisability of the results presented above.

References

1. Banerjee, S., Ramanathan, K., Gupta, A.: Clustering short texts using wikipedia. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 787–788. ACM (2007)
2. Banko, M., Brill, E.: Scaling to very very large corpora for natural language disambiguation. In: Proceedings of the 39th annual meeting on association for computational linguistics. pp. 26–33. Association for Computational Linguistics (2001)
3. Bollegala, D., Matsuo, Y., Ishizuka, M.: Measuring semantic similarity between words using web search engines. www 7, 757–766 (2007)
4. Caruana, R., Niculescu-Mizil, A.: An empirical comparison of supervised learning algorithms. In: Proceedings of the 23rd international conference on Machine learning. pp. 161–168. ACM (2006)
5. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning 20(3), 273–297 (1995)
6. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American society for information science 41(6), 391 (1990)
7. Gabrilovich, E., Markovitch, S.: Overcoming the brittleness bottleneck using wikipedia: Enhancing text categorization with encyclopedic knowledge. In: AAAI. vol. 6, pp. 1301–1306 (2006)
8. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: IJcAI. vol. 7, pp. 1606–1611 (2007)

9. Genc, Y., Sakamoto, Y., Nickerson, J.: Discovering context: classifying tweets through a semantic transform based on wikipedia. Foundations of augmented cognition. Directing the future of adaptive systems pp. 484–492 (2011)
10. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford 1(2009), 12 (2009)
11. Halevy, A., Norvig, P., Pereira, F.: The unreasonable effectiveness of data. IEEE Intelligent Systems 24(2), 8–12 (2009)
12. Holte, R.C.: Very simple classification rules perform well on most commonly used datasets. Machine learning 11(1), 63–90 (1993)
13. Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. Machine learning: ECML-98 pp. 137–142 (1998)
14. Kim, S.B., Han, K.S., Rim, H.C., Myaeng, S.H.: Some effective techniques for naive bayes text classification. IEEE transactions on knowledge and data engineering 18(11), 1457–1466 (2006)
15. Landauer, T.K., Foltz, P.W., Laham, D.: An introduction to latent semantic analysis. Discourse processes 25(2-3), 259–284 (1998)
16. Lewis, D.D.: Naive bayes at forty: The independence assumption in information retrieval. In: European conference on machine learning. pp. 4–15. Springer (1998)
17. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM 38(11), 39–41 (1995)
18. Rennie, J.D., Shih, L., Teevan, J., Karger, D.R.: Tackling the poor assumptions of naive bayes text classifiers. In: Proceedings of the 20th International Conference on Machine Learning (ICML-03). pp. 616–623 (2003)
19. Song, G., Ye, Y., Du, X., Huang, X., Bie, S.: Short text classification: A survey. Journal of Multimedia 9(5) (2014)
20. Sun, A.: Short text classification using very few words. In: Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. pp. 1145–1146. ACM (2012)
21. Wang, F., Wang, Z., Li, Z., Wen, J.R.: Concept-based short text classification and ranking. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. pp. 1069–1078. ACM (2014)
22. Wilcox, R.R., Keselman, H.: Modern robust data analysis methods: measures of central tendency. Psychological methods 8(3), 254 (2003)
23. Yuan, Q., Cong, G., Thalmann, N.M.: Enhancing naive bayes with various smoothing methods for short text classification. In: Proceedings of the 21st International Conference on World Wide Web. pp. 645–646. ACM (2012)